Congrès AFSP Toulouse 2007

« Table ronde 1 - Les méthodes en science politique des deux côtés de l'Atlantique »

Session 3

COLAZZO Dario
(LRI, CNRS UMR 8623),
dario.colazzo@lri.fr

DUDOUET François-Xavier
IRISES (UMR CNRS 7170)
dudouet@dauphine.fr

MANOLESCU Ioana
Projet GEMO, INRIA Futurs
ioana.manolescu@inria.fr

NGUYEN Benjamin
PRiSM (UMR CNRS 8144),
benjamin.nguyen@prism.uvsq.fr

SENELLART Pierre
Projet GEMO, INRIA Futurs
pierre@senellart.com

VION Antoine
CERI (UMR CNRS 7050) et LEST (UMR CNRS 6123)
antoine.vion@univmed.fr

Analysing web data-bases.
Towards new AI inquiries.

First results of the Webstand ANR project on W3C

ABSTRACT

This paper presents the first results of a study on the bargaining process of web standards in World Wide Web Consortium (W3C) arenas. This process is analysed through bargaining habits and through networks of actors who take part in it. The data collection is based on information about individuals who play the game: Who are they? In the name of whom do they talk? etc. Official chats and forums, web home pages and personal pages are exploited by ad hoc crawling methods, and stored in an XML data warehouse.

This warehouse is based on the semi-structured database model, which is a flexible tree-patterned schema, very different from the well-known classical relational tables and traditional data warehousing techniques. This structure can improve the management of sociological inquiries, by allowing an evolutive strategy of querying, based on categorization. It simplifies the research process prior to quantitative work (statistics, factorial analyses, structural network analyses).

In this paper, we present both the foundations and architecture of the warehouse, and the sociological results that have been gathered so far using these novel techniques.

KEY WORDS

Activism, Complex Queries, Key Actors, Key Institutions, Influence, Multi-level Analysis, Optimal matching methods, QCA, Semi-structured Data Bases, Social Networks, Web Warehousing, Web Data

1. INTRODUCTION

This original paper is joint work between sociologists and computer scientists, specializing in database technology. In order to help the non-expert reader, some computer science technical terms are explained in the glossary (section 5). When this is the case, the first occurrence of these terms will be in **bold font** *in the text. When the terms are important for the comprehension of the text, in order to simplify reading, this information may be found in footnotes.*

Web data is becoming a big challenge for scientists from many aspects: How can researchers follow the continuous variations of information and forums available on the web and build archives from them? How can they construct adequate peer-to-peer data bases to store and query these vast quantities of information? These problems are some

of the many recent concerns of *computer scientists*. From the *sociologists*' point of view, one may ask oneself related questions, such as how researchers should study these new ways of social communication and new kinds of governance that the web creates; all the more so by *using* these new technologies to analyse this data…know as *database technology*.

What is a **database**, and what is a **data warehouse**? We use both terms in this article. A database is simply a collection of data, stored on a computer, with various levels of security, granting simple access to the data stored. A data warehouse is built on top of a database, and provides more complex operations on the data. For instance, if we build a database on the key actors of the standardization process of the **World Wide Web Consortium (W3C)**, the database will provide access to their names, or messages posted, whilst our data warehouse will provide enrichment operations such as retrieving their CVs from the web, grouping the actors according to their institutions, etc. Devising a data warehouse is an extremely complex task, and is field dependent. Therefore in order to achieve the construction of a social science data warehouse, our research group comprises of both sociologists and computer scientists specialized in database management.

Originating from two fields, we are trying to bridge the gap between those approaches, in order to create and analyse new kinds of methodologies and studies. Our work consists in opening up paths to extract data automatically and build new kinds of sociology oriented, user friendly databases, which are no longer mere ***relational* databases**, but ***semi-structured XML***[1] **databases**, and to experiment this methodology on the case study of the standardization process of the web in the W3C, which is the first application we present among other ongoing studies.

In the following sub-sections, we begin with an introduction to the problematic of the analysis of Web data, and advocate the use of database technology to help sociologist with this task. In Section 2, we present our new methodology used to create and fill the data warehouse. In Section 3, we describe the first results of our study on the standardization process in the W3C. Section 4 is a conclusion.

1.1. Dealing with web data: methodological challenges and needs for new computational methods

A tremendous amount of information documenting various human activities from business to culture industry or information has moved online in the form of HTML, XML and PDF documents. For instance, national and international organizations are gradually publishing information online, which may change quickly and be very hard to capture without any automatic extraction process. A raw Google estimate of the number of Web pages under europa.eu.int (the European Union official website) is one million; those under gouv.fr (French government site) are estimated at 600 000 and those under .gov (US government sites including the Library of Congress and the National Institute

[1] The eXtensible Mark-up Language (XML) is a W3C standard for semi-structured data (i.e., documents). Along with the XPath, XSL, XQuery query languages, it is a standard for semi-structured databases just like SQL is a standard for relational databases. In our text, we use either XML or semi-structured in an equivalent way. We assume the reader is somewhat familiar with relational databases as described in (Codd, 1970).

of Health) at 15 600 000. While social science research could clearly benefit from Web data storage and analysis tools, these are currently not available to the sociologist. While some scientists do use general-purpose database management systems (DBMSs, most frequently relational DBMSs, such as Microsoft Access, IBM DB2, Oracle, etc.) to store their information, such data collections are most often done by manual insertion or copy-paste from a screen to a structural database format, since Web document formats do not fit the relational DBMS used. Such DBMSs are poorly suited to the social scientist's needs since they do not support the inherent heterogeneity of various sources (i.e., the different formats of representation of the information found in the *highly unstructured*[2] documents on the Web) and do not assist the scientist in the modelling and conception of these very specific applications.

The main contribution of our work is *a reflection* on how sociologists and computer science researchers can collaborate and produce specific application tools and methods of Web data analysis to complement the traditional interviews and statistical analysis. The second contribution of our work is *a generic computer application architecture* based on standards such as XML, XSL and most importantly XQuery for a Web data management tool to be used by social scientists collecting and analyzing Web-based data sources. Our approach innovates over the state of the art in sociology research by using Web technologies centered on XML and by providing database-style tools to analyze human interactions, as captured in mailing list content available on the Web.

1.2. Bypassing common sociological uses of data bases

When using databases, sociologists usually face two ranges of problems: the lack of control of their tool, and the technical limits which slow down the evolution of their work. The first range of problems is the lack of control, ranging from usual difficulties in the learning process to legal and intellectual problems which affect databases already constructed and filled. In this last case, working with available databases is of course time-consuming: Quite often, researchers are not allowed to enrich the data, and are seriously constrained by predefined categories (such as the **schema of the database**) they might not need or like. For example, studying childcare arrangements in France could lead to deal with lone parents as a category even though this category is not well defined by homogeneous criteria, and despite the fact that this category will not exist in older databases; therefore longitudinal approaches of childcare arrangements are difficult (Martin, Vion, 2001). Indeed, the main reason sociologists use such databases is the need to share time and their inability to construct a database from scratch. Of course, since database management has become a disciplinary field by itself, one should not ask sociologists to become experts in this field. But the main result is that few of them are able to elaborate a database that would be more sophisticated than a simple **table**. Nevertheless for a long time now, database managers have indicated better ways for enquiries, such as relational databases (Codd, 1970) or semi-structured databases (Abiteboul, 1997). Why not try to build a user-friendly methodology to guide the sociological use of semi-structured databases? The second range of problems emerges from some technological rigidities of database schemas. People having experimented the management of relational databases know that making their structure evolve as quickly as phenomena or their own perceptions is extremely difficult and challenging:

[2] Although defining structure information is very complex, it can be seen as pairs *element/value* such as *personName=Smith* indicating that a persons' name is « Smith ».

this is not the point of relational databases, whose structure is defined once and for all, at the moment the database is originally created. We feel and advocate that we should explore the opportunities given by a far more flexible schema such as a semi-structured one.

We present in this paper our first investigations about the ways sociologists could construct and use (XML) semi-structured databases in the future.

1.3. Case study: the standardization process of the Web

Our first field of experimentation is the sociological analysis of the establishment of a W3C Recommendation. The choice of this topic is due to our participation in a French government-sponsored project on the analysis of standardization processes in the area of Information Technology (IT)[3]; however we also envision a wide range of applications and possible extensions of this work. The study of information technology is an emerging hot topic for today's sociologists and remains a domain vastly unexplored. The success of innovative technologies depends on their widespread adoption which in turn depends on their recognition as a standard. Let us stress that recommendations such as those issued by the W3C, although not formal standards technically speaking are de facto standards once they are adopted by a worldwide user or industrial community (Dudouet, Mercier, Vion, 2006). Henceforth we will use the term standard to refer to such technical documents. The role of the sociologist in standard-setting bodies such as the W3C is obvious: such bodies are concerned about the usability and accessibility of Web technologies to the greatest possible number. Understanding the processes of communication, technical initiatives and standard production is useful for any organization (academic or corporation) with a focus on Web technology. Such an understanding is equally useful for the standardization body itself as it can lead to improved or better explained procedures. The need for a social analysis of IT, however, goes well beyond standard-setting bodies; even major companies such as Microsoft are hiring sociologists to analyze interactions on UseNet message boards (although their numbers are still thin for now).

Abbott and Gilbert (2005) recently invited scholars to "use computational methods as a means to an end—the advancement of sociology—rather than stopping short when they have developed a model that *works*". This is why we first present our general intentions for advancing sociology by evolutive and flexible data bases (Section 2) before presenting what "*works*" in our ongoing study on the bargaining of web standards (Section 3).

2. EVOLUTIVE AND FLEXIBLE DATA BASES: METHODOLOGICAL STAKES FOR SOCIOLOGY

2.1. Conceptual modelling and schema of databases

[3] *WebStand Project,* including ACI Normes, Régulations et Pratiques des Politiques Publiques 2004 and ANR Jeunes chercheurs 2005.

The **schema**[4] of data warehouses depends on the sociologists' categories and on their hypotheses. Databases are constituted by fundamental abstract elements which are as many categories as needed and have a logical signification: **entities, attributes, relationships**[5]. These elements are a specific characteristic of the programming language but have no empirical meaning. The meaning of the data is constructed by the sociologist at two key moments: when conceiving the schema of the warehouse, and when interpreting the data.

If a database concerning people interacting on forums (see 3.) is made up without defining entities such as [mail] or [individual], it will be simply impossible to measure anything about individuals who posted mails. The choice of categories, which means a sharp definition of them, is not only crucial for the achievement of the study, but for the construction of the research topic. These preliminary remarks seem trivial, but represent a big challenge when trying to conceive a good schema. When one constructs an entity, one postulates that its essential characteristics will be informed in the database (i.e., the @name of a [person]). Some other characteristics may not be, but they will become secondary ones, so that they will not help establishing relations between entities of the base. For example, if one works on professional trajectories, at least two entities will be needed: [individual] and [firm]. The [individual] entity will be defined by primary characteristics, such as @firstname, @lastname, and many secondary characteristics such as @sex, @age, @status, @diploma, @skills, etc. The link with the firm will only be managed via a unique characteristic of the individual called **primary key**(either the name, if it is unambiguous or some sort of code), which means no anonymous individual could take place in cross-company comparisons, except if complementary models are added to bypass the lack of information (Jansen et al., 2006).

Anyway, correctly defining entities and their relations within a warehouse is a big challenge, and it is far from being meaningless for social scientists. Would all sociologists give the same definition of *individuals*? Of course not, and Weber's nominalism is very helpful to understand that such notions are constructions elaborated to draw up meaningful patterns of thought, based on *idealtypes*. Though it is strongly affected by such epistemological problems throughout the implementation of the programming process, database management science is somehow neutral from this perspective. The global pattern it offers is a kind of meta-theoretical one, as the ontological status of entities has no serious methodological implications on the construction of the database. For example, like Latour and Callon (1992), one could consider objects rather than individuals, and specify which of them are human and which of them are not.

Elaborating such a database forces sociologists to sharpen the notions they use in their models. Therefore this neutrality is all the more helpful since it takes into account as many "ways of world-making" (Goodman, 1978) as people would like to. This is why the method we experiment is not fully embedded in our study, but is virtually a generic know-how for any sociologist confronted to massive web data.

[4] The schema describes the structure of the information we want to store in the database. The more generic term *architecture* can be used, if we want to describe both the logical (schema) and physical (i.e., extraction modules, programs, disk partitioning, etc.) aspects of a database. In this paper, we focus on the schema aspects.

[5] In the text we represent entities in between brackets i.e., [person] ; we represent attributes prefixed by @ i.e., @name ; we represent relationships as lines between entities i.e., [person]--<posts>--[mail].

Any social scientist who wishes to construct a database schema has to elaborate a conceptual sociological model that fits with database management science. We feel that this task is achievable by the sociologist. For example, about professional trajectories, one could elaborate a simple model such as:

Entities:
[individual], [institution], [time period]
Relational scheme:
[individual]--<function within>--[institution]--<during>--[time period]

Unfortunately, translating such a conceptual model into a database schema requires skills that few sociologists have already acquired: They need help from computer scientists. Our ambition is to propose a user-friendly general model of conception, which would take into account the extended sociological world-makings (from entities such as individual, institutions, etc.). This would be composed of bricks that researchers could modify at will. The use of the semi-structured approach is compulsory here, due to the technical limits of relational databases.

2.2. Technical limits of relational databases for social studies

Sociologists have to face two difficulties when using relational databases: taking into account the evolution of the observed phenomena, and managing the evolution of their hypotheses. These two aspects generate the same problem from the perspective of the database management. Sociologists know well that they are often led to modify their point of view, adapt their hypotheses or simply deal with unpredicted phenomena. In our example of professional trajectories' study, one could wish to know whether the growth of *functions* and *wages* is quicker when people leave one firm and join another one. From the database perspective, the notion of function is not independent. If it is taken into account, it may appear as an attribute of the entities [firm] or [individual] or as an independent entity. If it is an attribute of [firm], this means only one single @function by [firm] is achievable (at a time). If it is an attribute of [individual], this means only one single @function by [individual] (career) is achievable (at a time). If one would like to know a little bit more about cumulating functions or their evolution, they must create an entity [function] and an entity [time] and develop the following relational scheme: [individual]--<is related to>--[institution]--<by>--[function]--<held in>--[sequence] (time interval).

However, with relational databases, changing attributes into entities is sometimes impossible, and if it is, the programming work is heavy and often data can be lost. Such problems disappear with tree-patterned semi-structured databases, which allow any modification of the structure without reconstructing the whole schema and losing data. The method we are experimenting is dynamic, because warehousing is evolutive and simplifies the integration of new hypotheses and phenomena.

2.3. Managing semi-structured XML data bases

As soon as the set of fundamental entities for sociological studies is stabilized, one major problem, as we showed, is to deal with the perpetually changing attributes of these entities or relations between these entities.

In contrast with the fixed structure of a relational database, semi-structured databases allow the dynamic addition of as many entities and attributes as needed. In this way, researchers will be able to preserve *both the global quality of the structure and the appropriateness of details they would like to keep available*. From this perspective, the particularity of sociological inquiries needs to be seriously taken into account when conceiving warehouses of sociological data.

In computer science, there has already been much work on the topics of data warehousing mediation and integration of data (Widom, 1995). We refer to (Chaudhuri, Dayal, 1997) and (Vaisman, 1998) for a survey on **On-Line Analytical Processing** (OLAP), data warehousing and materialized views and of methodologies and tools for constructing classical data warehouses. Standard OLAP and warehouse technologies only deal with highly quantifiable data.

In the case of the construction of a sociological warehouse the approach is radically different: We need to go from the notion of data warehouse to that of ***content* warehouse**. Content warehousing, which has been introduced in (Abiteboul, 2003) and (Abiteboul *et al.*, 2002), amounts to considering a warehouse of qualitative information (such as sociological data), that has no trivial mathematical processing method. This is needed because relations between participants in a W3C arena, or their exact roles inside this arena, do not lead to information that can be processed in a regular OLAP approach. This qualitative information is often available from various sources on the web is highly heterogeneous and can only be integrated by using a flexible semi-structured data model. We developed in (Abiteboul *et al.*, 2002) a methodology for the design and construction of a content warehouse focused on food risk: The e.dot (e.dot, web site) project based on the Active XML development platform (ActiveXML, web site).

Data integration and XML Data integration systems offer the possibility to query heterogeneous and distributed data sources as if they were contained in a single database. The classical architecture of a data integration system includes a mediator (Wiederhold, 1992) offering the integrated database view to the user and a set of wrappers which make the connection between individual data sources and the mediator. Research in this area has produced several data integration prototypes, among them we cite Disco (Tomasic *et al.*, 1996) and Garlic (Haas *et al.*, 1997) some of which have been transferred into industrial products (respectively the KelKoo comparative shopping portal (Kelkoo, web site) and the IBM DB2 product suite). Such systems could solve some of our problems but they have as such several disadvantages First writing wrappers is a very tedious and error-prone task; in contrast we take advantage of accepted standard Web data formats (in particular XML) to automate as much as possible the process of data extraction (see 3.2 below). Secondly they are oriented mainly on relational data while the very nature of Web documents requires handling XML. Thirdly such systems are typically query-intensive: They focus on the efficient execution of complex queries on large distributed data sources and give little or no support for the modelling stage where the concepts relevant to the sociological study are

defined; this stage is crucial in order to make the data analysis queries relevant for the social study.

As a summary, we have chosen to represent our content warehouse in XML for the following reasons:

- XML semi-structured format allows modification of the warehouse's schema and architecture in a large extent.
- XML represents semi-structured information in which structured data (e.g., for each message, e-mails and dates) can be mixed with raw text (message body), images, and so on.
- XML is flexible and evolutive: new information can be added at will by adding new elements or attributes.
- XML is intended to be the language of the Web making it suitable to the writing of wrappers for Web pages or other data found on the Web.
- A mailing list has an inherent tree structure (message A is the child of message B if B answers to A) which requires a nested representation format such as XML. As soon as one wants to extract and store such data, the more isomorphic their database would be, the easier queries would be achieved.
- XML remains simple to understand. Sociologists can grasp the extent and the kind of the information stored in the warehouses just by reading the XML data.

Therefore, for this and many similar applications, XML is a real step forward for quantitative sociological analysis which has traditionally been carried out in the context of older databases. The way such warehouses fit with available software methods for sociological analysis is very open. Sociologists today are dealing with a large scope of information (interview, archives, mailing list, statements, etc…) in various format (paper, audio, video, electronic format, themselves divided in text, PDF, HTML, XML, etc., formats). Data warehouses have to challenge this variety of data to permit development of sociologists' enquiries in the largest extent, so as to conquer new realm of investigation such as the Web. Creating a generic platform in XML, which can integrate documents of various formats and from various sources, will allow to export our data to many software conceived for social sciences (SAS[6], fs/QCA[7], Pajek[8], TDA[9], etc.).

---

[6] SAS is the most generally used software for statistical analysis, as it allows to implement any kind of measure, from simple factorial analysis to linear regressions, probit, logit, etc. All our data can be exported to SAS.

[7] QCA and fs/QCA use a relational structure but our data can be exported into tables. As Caren and Panofski (2005) also pointed out, another constraint comes from the fact QCA makes an extensive use of boolean mathematics, which limits the heterogeneity of data. Our data might be richer and too heterogeneous for QCA (see 1.1). Another aspect that could be inspired by QCA could be to integrate QCA methods in XQuery in order to add functionalities.

[8] Pajek is the most advanced network analysis software. It is used in its version 1.18 in our study (see below in 3.3).

[9] TDA is the most advanced software allowing to manage *optimal matching methods*. Optimal matching methods (Abbott, Hrycak, 1990; Abbott, 1995; Abbott, Barman, 1997), are a very simple way to distinguish regularities through and within diachronic sequenced data. This method is of particular interest for the study of multi linear trajectories such as careers (Stovel, Savage, Bearman, 1996) or sequentially organized cultural artefacts (Abbott, Tsay, 2000) or social habits (Lesnard, Saint-Pol, 2004). Even though we have not yet experienced such a method, we estimate it as theoretically achievable from XML data bases. TDA is available at http://steinhaus.stat.ruhr-uni-bochum.de/binaries.html, the interface Win TDA at: http://www.tufts.edu/~kschmi04/research/ and the manual at http://www.stat.ruhr-uni-bochum.de/tman.html

3. APPLICATIONS: FIRST EXPERIMENTATIONS ON THE STANDARDIZATION PROCESSES WITHIN THE W3C

Our study on the W3C is part of a larger study of the international standardization process. Based on previous work, in particular on international drug control policy, we investigate how standards are established at an international level. These standards end up by governing the practices of billions of individuals[10] in specific domains.

3.1. Hypotheses and processes

The main hypotheses of our study are the following ones: 1) Standards are elaborated by small networks of experts who developed common and rare know-how 2) Standards structure new markets so that standardization processes represent economic stakes for firms 3) Standardization processes consist of letting a group of people or firms to monopolize formats for industrial applications (Dudouet, Mercier, Vion 2006). These hypotheses are tested in two fields: mobile phone technologies and web technologies. We present here a part of our study on web standards.

In order to test our hypotheses on standardization processes of the web, we decided to manage a kind of multi-level analysis[11] by observing the concrete activity of experts of the W3C and then to look for their ties with companies or any other institutions concerned with innovation. Such a method leads to deeper analyses and systematic measures of the structural dynamics of the standardization work investigated by Tamm-Hallström (2001, 2004) or Graz (2006).

The activity of experts mainly consists on arguing and bargaining on mailing lists in which recommendations for standards are debated. These recommendations become statutes for programmers. The W3C public working group pages capture communication interaction between different people (called *actors* in the following, and modelled as entities) and trace their actions, positions and declarations through time. These Web-based information sources inform the sociologists about the interactions between actors of a given process[12]. In the particular case of elaborating web standards, mailing lists are all the more interesting since they are becoming the prevalent means of interaction between participants scattered around the globe and working in different time zones. Moreover when participants do actually meet physically or attend teleconferences, written notes of the meeting or teleconference are taken and usually published on the Web shortly afterwards. A social scientist studying the standardization process must adapt to this situation by developing and using computerized data

[10] Acording to the web site http://www.internetworldstats.com/ there are over 1.1 billion internauts in 2007.

[11] Hox and Kreft (1994) proposed a smart survey about multi-level analysis, and the use of such methods in sociological inquiries. The main problem is of course managing the variation of contexts when jumping from individual behaviours to - for example - institutional patterns. From this perspective, see our first results below.

[12] A few available studies have explored the social uses of mailing lists from the perspective of posting habits (Buckner and Gillham, 1999; Dudouet et al., 2005), sociality rules (Beaudouin, Velkovska, 1999), or network structure and ethnology (Auray, Dorat, Conein, Latapy, 2006).

management tools since techniques such as manual interviews and information collection become powerless to apprehend the sheer size of the corpus.

Our research process aims at investigating the structure of interactions on mailing lists, in order to understand who leads the discussions, both qualitatively (pragmatic analysis) and quantitatively (network analysis). Both aspects need to be related. Quantitative analysis mainly consists on counting the numbers of e-mails sent on mailing-lists by individual actors, actors from a given institution, etc. This gives first indications on the level of personal investments in this activity. For this paper, we present the quantitative work, and how we went back from mailers to institutions (firms, research centres, NGOs, etc.) the actors were linked to. Our aim was to identify which institutions were the most implicated in the standardization process. Since we had to deal with about 20.000 mails and around 3.000 email authors, no manual work could be done. Therefore we built a specific conceptual model, which may have further applications for various other sociological enquiries based on large groups of people communicating via emails or forums.

In this specific model, entities are [actor] (individual), [institution], [message], [arena], [function], and [time]. The model captures the fact that [actor]s have one or several e-mail addresses from which they send [message]s. These actors hold functions in [arena]s (mailing lists, working groups, commissions, departments) which represent a division of an [institution]. For example, a developer, John Doe from Microsoft who sends the message "Hi folks, I'm new to this list" to the mailing-list public-qt-wg of the W3C XQuery working group is conceptualized as followed:

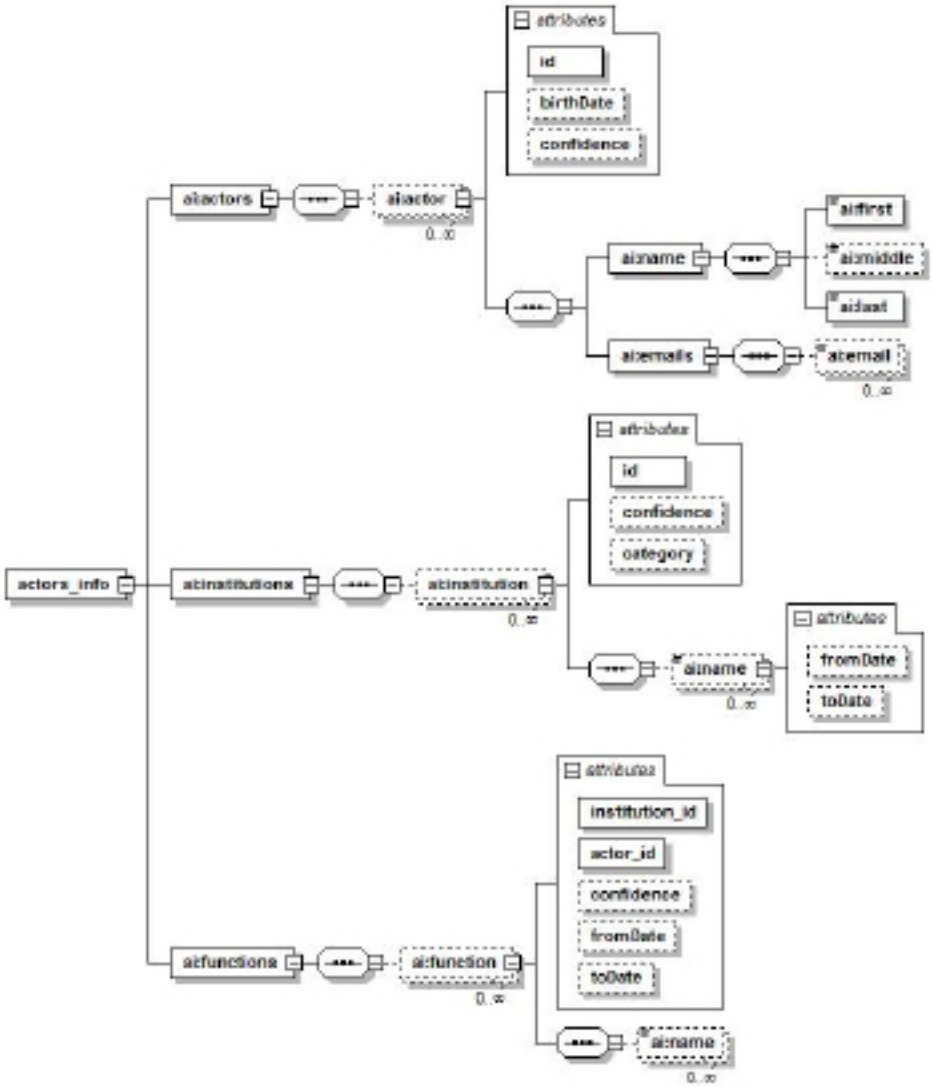
> *John Doe* [name] is an *employee* [function] in a *unit* [arena] of *Microsoft* [institution] who sends *"Hi folks, I'm new to this list"* [message] at a specific moment [time] on the mailing-list *public-qt-wg* [arena] of the working group *XQuery WG* [arena] of the *W3C* [institution] and who is *de facto* becoming a *participant* [function] of the mailing-list *pulic-qt-wg* [arena].

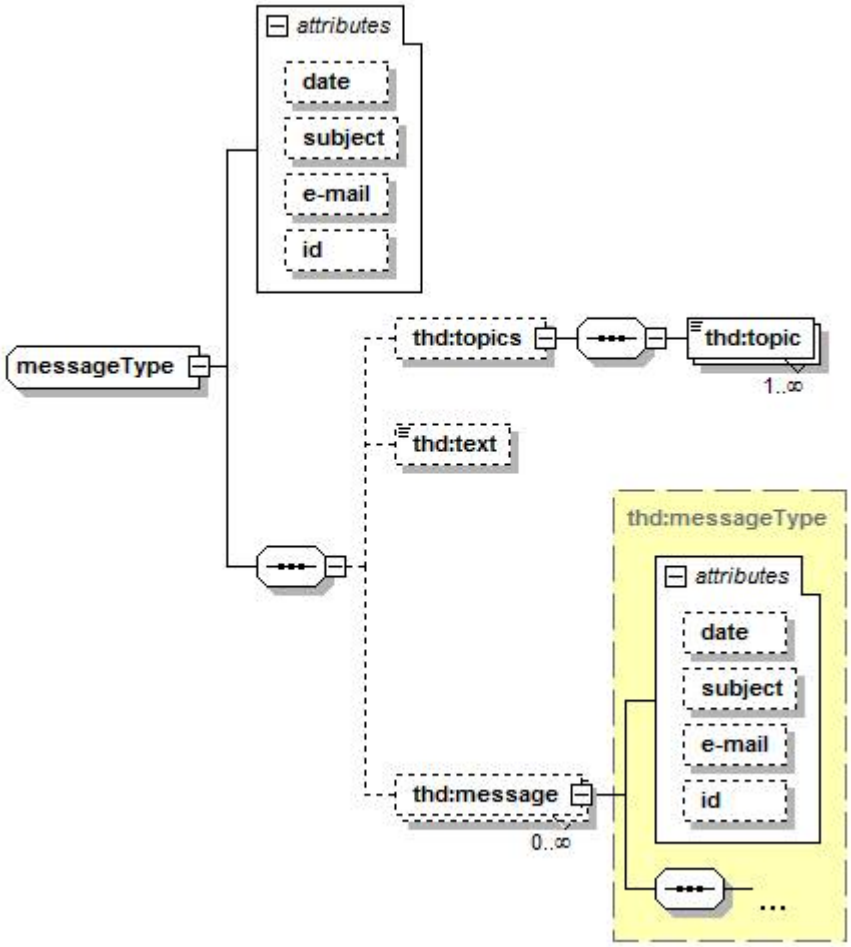
The other side of our model deals with the representation of messages of the mailing list; arguably the most important information retained here are the links between a message and its author (an [actor]), the **tree-structured** *threading* of the messages (which message answers to which), as well as the full textual content of the e-mail. Other information available is the date, the subject, as well as the identified topics of the message.

The following diagrams present these two models in more detail. Let us give a few comments on their structure, although we do not want to enter into the details of the representation. These diagrams are graphical representations of the XML Schema of the database. They define elements and attributes, linked together in a hierarchical manner. Some elements and attributes (dotted lines) are optional, while others (full lines) are compulsory. Cardinality of the links (i.e., the number of nodes that can be under a given element) are also indicated. For instance, the 0..infinity cardinality under [actor] means that there can be any number of [actor] under the [actors] element.

The second schema shows the recursive tree structure used to store the messages. The top level element is a [thread] that can have any number of [message] elements

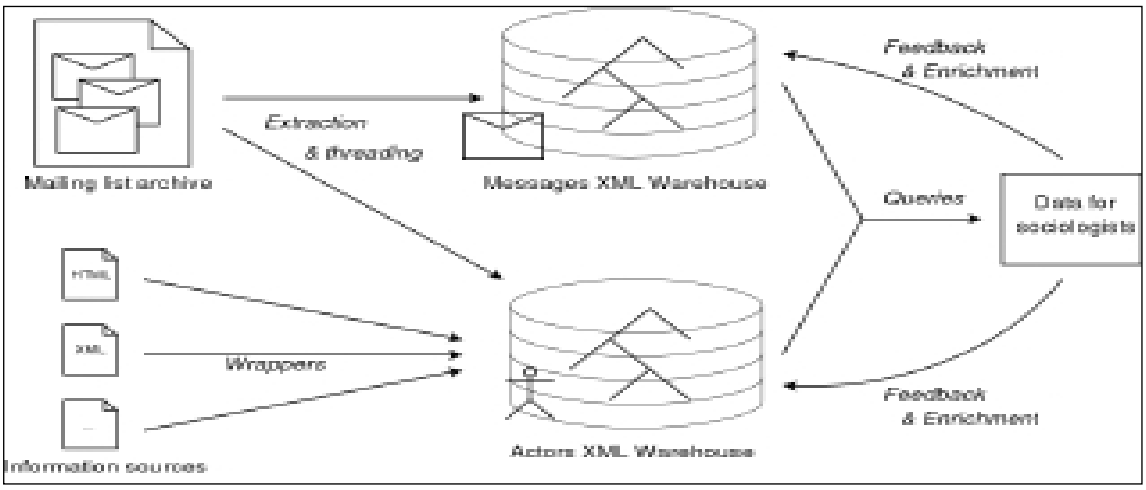
underneath it. Each message may in turn have as many other [messages] that answer to it.





3.2. Managing Data: Automatic extraction / cleaning / manual validation / enrichment

In order to exploit very large quantities of information, it is critical to design a methodology which needs as little human intervention as possible. However human input and feedback can (and should) be used to tune and enrich the system. For example, when extracting data from mailing lists, our process model is organized as follows:



As shown on our schema, we are interested in identifying actors: the individuals that post messages on the mailing list. Each author has a unique ID and first, middle and last name. An actor can have multiple e-mail addresses. Furthermore an actor can have multiple roles within different institutions i.e., be both a university professor and a consultant for a company. Messages are posted from an e-mail address. We capture the date, author, subject and text of each message. This model is the starting point of our analysis. Once the conceptual model is established (see 3.1), we map data sources of interest to entities and relationships of this model and load the data sources into our warehouse.

A number of practical issues require the use of automatic and semi-automatic filtering and enrichment: The name of an institution may appear written in many different ways. We devise generic patterns to recognize that “Sun Microsystems Inc.” is the same institution as “Sun” and “Massachusetts Institute of Technology (MIT)” is the same as “MIT” Further identification can be done manually (i.e., to understand that “cerisent.com” is a Web site belonging to “Mark Logic Corporation”) A person can have several different e-mail addresses; identification of persons must thus be done on their names and not on their addresses (assuming that two persons do not have the same name). In a relatively small group (involving about 200 people) a simple manual check, based on automatic extraction of names from email addresses suffices to make sure two actors do not share the same pair firstname/lastname (which we found to be indeed the case in our study). For a larger group standard data cleaning tool on a person databases could be applied. Currently, there is some research going on in the database community to find simple ways of defining if two homonyms are or are not the same person, based in general on their links to other entities (Kalashnikov, *et al.*, 2007).

In our case, first names, optional middle name, and last name are extracted from the full name description making possible the assimilation of “FirstName LastName” and “LastName FirstName” for instance. We have not yet introduced any data cleaning based on structural relations.

The content of these two warehouses is what we call “raw” data; we may enrich it manually with extra information on actors and institutions. The data extracted automatically can be complemented by other information sources found for instance on the World Wide Web (HTML or XML data describing mailing list posters, home pages of the more important actors institutions websites) using wrappers. Our investigations now aim at using information retrieval processes in order to ease such work. For instance, to efficiently find Web pages of an actor or organization, we can exploit some techniques aiming at identifying relevant pages by using some non-content features, like page length and the URL form (Kraaij, Westerveld, Hiemstra, 2002). In addition, we can exploit other techniques. These are based on the use of existing web search tools (e.g., Google) to find a coarse list of potentially relevant pages, and on the use of particular information retrieval techniques able to extract relevant and representative information contained in these pages, to guide the user in the identification of searched Web pages (White, Jose, Ruthven, 2001). It is of course technically possible to combine automatic fetching and information retrieval from available data. In any case, our procedures will always be semi-automatic, which means we do not aim at creating

humanoid robots such as automatic *sociologists* or so on[13]. Our system generates propositions which need to be manually confirmed.

3.3. Measuring activism and influence: from key individuals to key institutions

The data we present is not the final results of the study. Indeed, temporal variations are not sufficiently taken into account at this step of the process. As our results are still too synchronic, we prefer talking about first observations that guide the future developments of the research process.

The corpus is composed of 8 public mailing-lists of the W3C, which have been active from 1999 to 2006 and we extracted automatically and fully. In addition, we have made an inventory of the technical preconizations[14] produced in connection with these lists. The chosen lists were about the XML standard, as it allowed us to have a good control and knowledge of the data for a first experimentation. The first range of investigations consisted on measuring activism on these lists. Note that these results are a continuation of our work exposed in (Dudouet *et al.*, 2005), where we focused exclusively on the public-qt-wg mailing list. We considered as active the participants who posted at least 20 messages on a mailing-list. We obtained 72 actors who sent 10619 messages on 8 lists, which represents 61 % of the total interactions if we exclude all the purely administrative mails of the W3C management team (3944 messages).

In order to show the variety of participants (in terms of posted mails), the relative importance of the mailing-lists and those of the participants who were active on several lists, we made the following graph (Graph 1). We represent mailing lists as green diamonds, and actors as red circles. The larger the size is, the more important the poster or list is, in terms of number of messages. We have been force to reduce the size of the graph to show its global properties. A larger version of the graph, or zooms on subparts are available on demand.

[13] For a survey of such humanoid social robots, see (Zhao, 2006). In France, the best robot available for sociological longitudinal qualitative studies is Marlowe (Chateauraynaud, 2003), elaborated by the Groupe de Sociologie Pragmatique (GSPR), a research center of the Paris EHESS.

[14] By technical preconizations, we include: official recommendations of W3C, drafts, which are supposed to become such recommendations, and Working Group notes.

Graph 1. Mapping[15] of the activism of individuals on the public mailing-lists of the W3C concerning XML standards (using Pajek 1.18)

The first observation we can make is that mailing-lists are not equivalent in terms of flow (number of messages), but they are all linked by at least two actors. These key actors (14 among 72) are multi-posters. Activism, in terms of posted messages, is of course mainly due to these types of participants. But we also find sometimes activists on single list. This analysis is not sufficient to lead to anything conclusive, but remains very useful to map the main actors of the whole network and to make hypotheses on their social influence (Marsden, Friedkin, 1993)[16]. Why do these people invest so much

[15] We would not talk about a social network here, as we are aware of the fact the non inclusion of posters of less than 20 messages in the one or two shorter lists (less than 500 mails for a whole) can represent a bias. As Kossinets (2006) recently showed, fixed choice designs can dramatically alter estimates of network-level statistics. But the author also finds that social networks with multiple interaction contexts may have certain interesting properties due to the presence of overlapping cliques. Our mapping aims at revealing the main individuals of such cliques without measuring the whole interactions within the public lists.

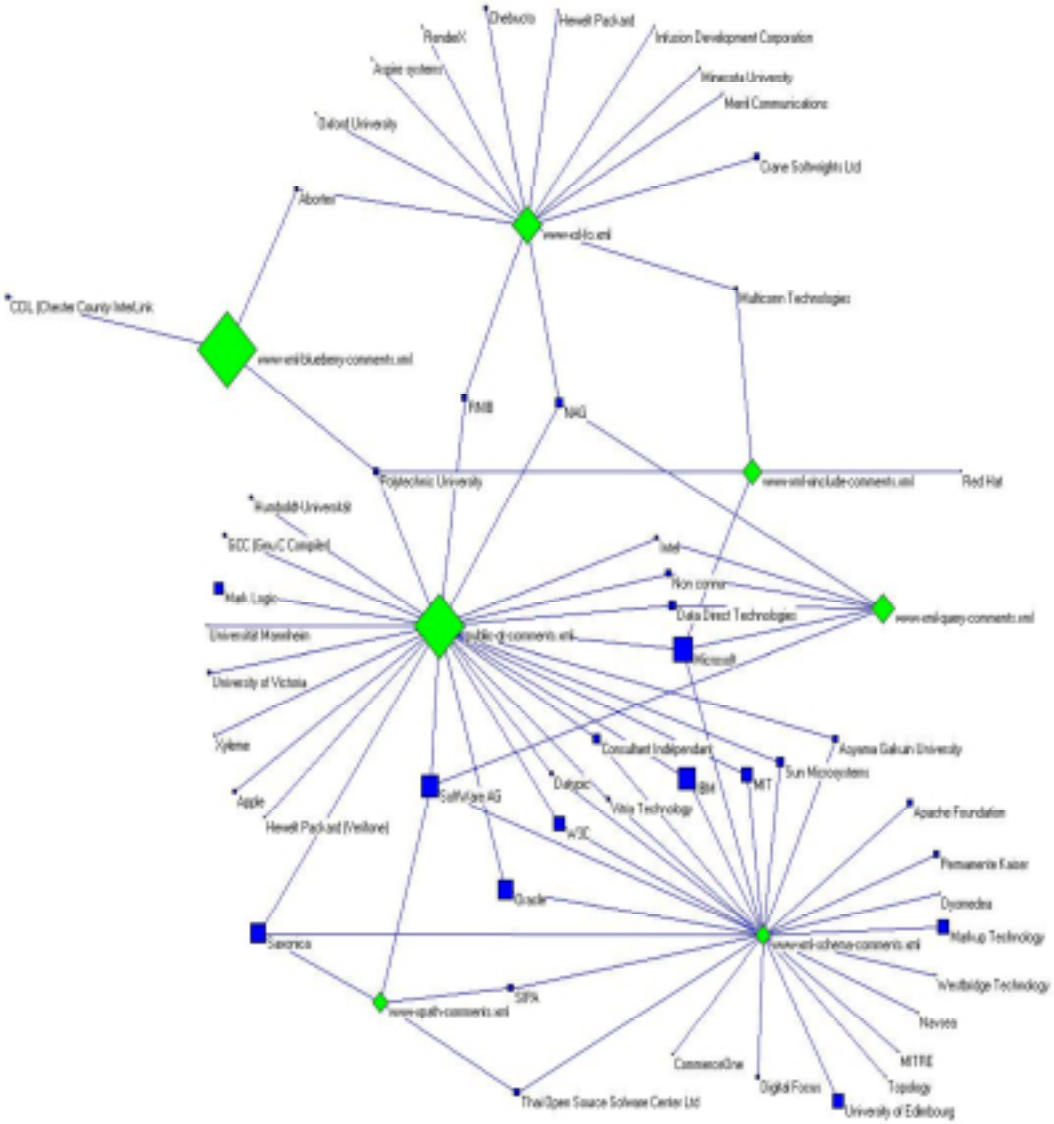
[16] We of course make a conceptual difference between activism and influence. The only author drawn in red on the mapping, Tim Berners-Lee, invented HTML and founded the web. He is so influent that he is called 'God' by some computer scientists. In our public lists, he only posted one mail. We suppose that his case is somewhat exceptional in our corpus.

time in these mailing-lists, and does their activism have any consequence on standardization process? We now have to go back to the recommendations to check the correlation between activism and influence on the final standard (see below).

A promising result of our methodology is that it captures all key institutions involved in the standardization process of the web in a vivid picture. Sociological studies inspired by new institutionalism have been concerned with the ways firms build durable networks and use them to buffer uncertainty, hide or restructure assets, or gain knowledge and legitimacy, by building durable networks (Stark, Vedres, 2006). Powell et al. (2006) are also concerned with innovation in their study about the development of biotechnology. By applying analyses of the structure and dynamics of the networks, and building statistical models to describe their growth through a period of eleven years, they show how the linkages evolve and how this is related to the changing involvement of institutions: universities, research institutes, venture capital, and large and small firms. At the moment, our study is less longitudinal than Powell et al's one. Nevertheless we can give some interesting observations with the work done so far. Observe Graph 2 (on the following page). Green diamonds are once again mailing lists, and blue squares represent institutions. The larger the diamond or square, the more messages posted.

As the graph shows, firms are very active: people from Microsoft, Software AG, IBM, Oracle, Saxonica, have sent more than 1000 messages. To be very rigorous, one has to be cautious with this measure, because some individuals may have had successive memberships fuddled by the synchronicity of the data. Though all institutions are affected by this problem, private ones might be more than public ones. Anyway, mostly represented institutions (those connected with at least 2 mailing-lists) on public mailing-lists are firms: 14 firms in comparison with 8 of other kinds (research centres, associations, NGOs). Even if our measure has to be sharpened by a more longitudinal approach, we can say that the biggest US AI companies play a big role in these mailing-lists. One can note the absence of Google, as well as the quasi-absence of non US firms, if we except Software AG. Research institutes and University play a secondary role.

In order to understand concretely the standardization process, we next relate our study of activism on these lists to the study of all technical preconizations produced by experts about XML standards, in Table 1.



Graph 2. Activism of institutions[17] on the public mailing-lists of the W3C concerning XML standards (using Pajek 1.18)

[17] Membership of institutions has been here manually reconstituted.

Institution	Type of Institution	# Individuals	Total Text	W3C Recommendations	W3C WG Notes	Working Drafts	Activism	Rank of activism (on 51)
IBM	Corp	11	13	8	2	3	1401	3
Oracle	Corp	8	13	6	1	6	1289	4
AT&T	Corp	2	7	4		3	_	_
Data Direct Technologies	Corp	1	6	2	2	2	363	11
Microsoft	Corp	5	6	4		2	1780	1
BEA Systems	Corp	2	3			3	_	_
Infonyte GmbH	Corp	1	3	1	2		_	_
Library of Congress	Gov	1	3			3	_	_
Unknown	n.a.	2	3	3			_	_
Sun Microsystems	Corp	1	3	3			300	12
University of Edimbourg	Uni	2	3	2	1		591	7
Mark Logic	Corp	2	2			2	418	10
Saxonica	Corp	1	2	2			1062	5
University of Venice	Uni	1	2		2		_	_
Brown University	Uni	1	1	1			_	_
CommerceOne	Corp	1	1	1			42	37
INRIA	Uni	1	1			1	_	_
Inso	Corp	1	1	1			_	_
Invited Expert	n.a.	2	1			1	_	_
Kaiser Permanente	Org	1	1	1			_	_
MIT	Uni	1	1		1		571	8
Pisa University	Uni	1	1			1	_	_
SIAC	Corp	1	1	1			_	_
W3C	Org	1	1		1		615	6
WebMethods	Corp	1	1			1	_	_

Table 1. Test of the correlation between activism and influence on final versions of official texts

First, the table shows that the correlation is not absolute. Some of the institutions associated to technical preconizations do not appear from our measure of activism. This can be explained by two main ways:

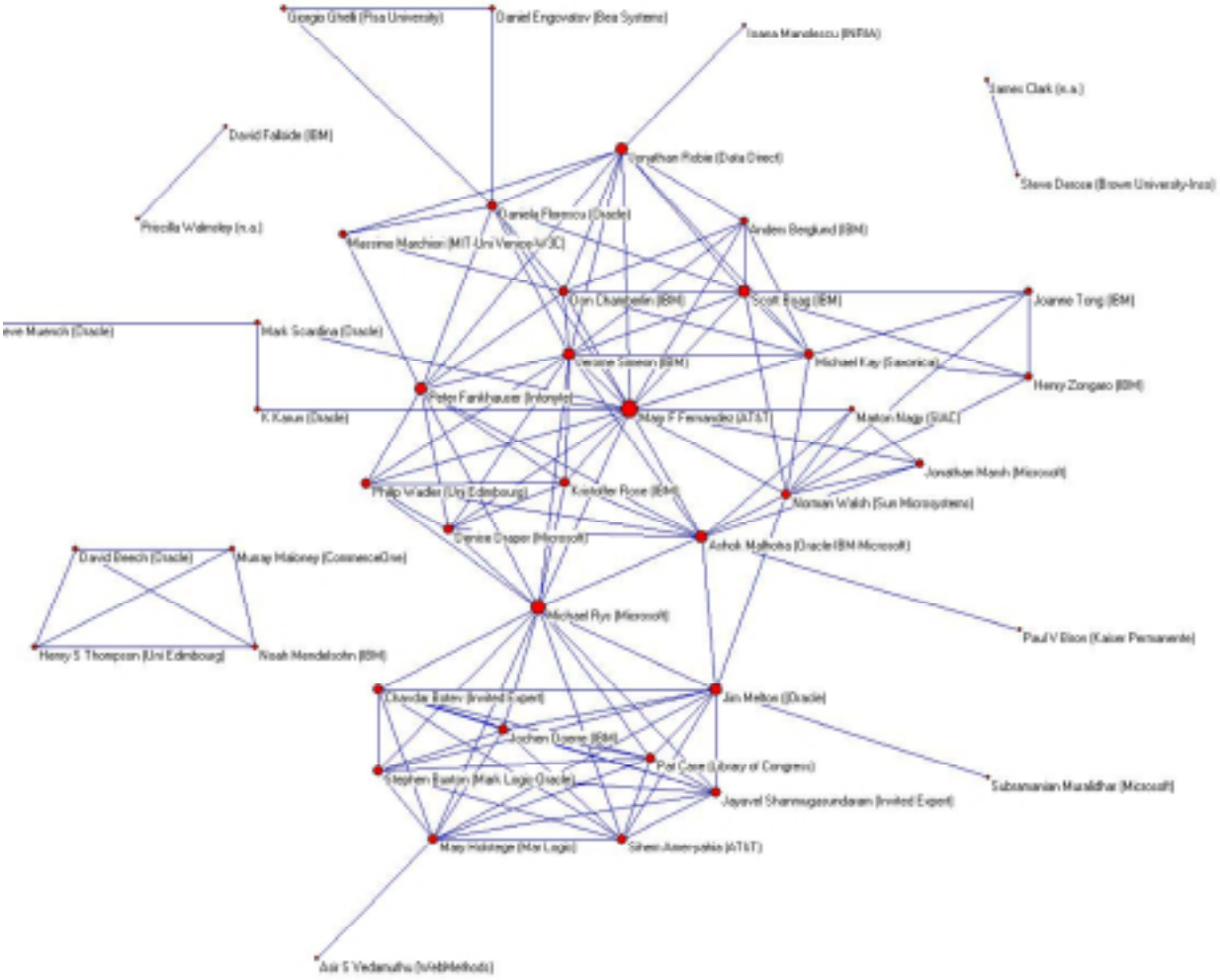
- the lists we used are public, but in some cases preconizations are decided in private lists for WG members only. As we cannot legally access to these private lists, some individuals and institutions are not taken into account[18]
- some of the individuals actors we take into account hold multiple functions. For example, Michael Kay (Software AG and Saxonica) only publish one membership (Saxonica) when authoring official preconizations.

In any case, one can observe that the ten most active institutions directly take part in the writing of final texts, except Markup Technology and Software AG (if we do not take Michael Kay into account). On 28 technical preconizations, firms (blue boxes) are associated to 26 in comparison with non-benefit institutions (cyan boxes): 8 academics, 4 unknown, 3 governments, 2 associations. This could be due to the human capital firms are able to mobilize in such processes. From this perspective, the institutional mapping of the authors of recommendations we take into account (graph 3) and the structural network of co-authoring (graph 4) lead us to build new hypotheses for more qualitative research.

Graph 3. Institutional official mapping[19] of the authors of recommendations

[18] W3C now incites actors to debate on public lists, what will ease our future studies.

[19] Memberships are here automatically selected and the data only take into account the ones declared by authors on the formal recommendations they sign. We see for example that Michael Kay only appears as a member of Saxonica. We will talk of an institutional network as early as we have checked all these missing data.



Graph 4. Structural network of co-authoring

When comparing graphs 3 and 4, we see for example that an author like Ashok Malhotra is a big poster; that he links together the main institutions, and is co-author of many recommendations. But we also observe that Mary Fernandez, who seems not to have big institutional resource from our official mapping (graph 3), manages to be one of the main co-authors (graph 4). This leads us to make more qualitative hypotheses about her trajectory, her personal networks, a strategic recruitment of AT&T, and so on. Graph 4 shows the firms hegemony in the network of technical preconization editors. All are represented at the centre of the configuration.

This finally means that we do not take this data as a means at the end, but we only stress that such data would not have been built up from any manual research. This encourages us to go further, and to experiment all the techniques we have in mind. For example, once we have grasped all these people CVs by information retrieval techniques and analyzed their trajectories by both longitudinal analyses of their co-authoring networks and optimal matching of their memberships, we will have complete data on people who make XML standards. This process could be extended to all the W3C public-lists in the future, what would mean we would have a very large understanding of the standardization process of the web.

4. CONCLUSION

In this article, we have shown how the use of state of the art developments in database technology can lead sociologists to be able to apprehend vast quantities of data readily available on the Web. We advocate that database modelling techniques, schema conception can help sociologists build and exploit large quantities of data, while also helping them to devise their model of reality. We also have shown that a preliminary analysis of mailing lists, and recommendations leads to the confirmation of the hypothesis that standards are elaborated by a small network of experts. Our analysis has led us to narrow our hypothesis in so far as we now see that industrial companies dominate the making of standards. We expect to investigate this further, using analogous techniques.

5. GLOSSARY

Database :

A database is a computerized system whose goal is to manage unlimited quantities of data. The term is quite vague, and can be applied to files, relational databases, XML Databases, depending on the type of data stored.

Database (relational):

Traditionally, since Codd (1970) one refers to Databases as *Relational, or SQL* (Structured Query Language) databases. Databases most used in the world are Oracle, IBM DB2, Microsoft SQL Server, Microsoft Access, or the freeware MySQL. Databases offer powerful querying facilities, implemented by the query language (SQL, XQuery).

Database (XML/Semistructured):

Semi-Structured or Native XML Databases are still hot research topics in Database Technology. The difference between a relational and an XML Database is the format of information that can be stored. In an XML Database, information is far more flexible, which leads us to prefer this type of data store for sociological applications, where the schema is difficult to define once and for all.

Data Warehouse:

A Data Warehouse generally speaking is a large commercial database, with added functionalities geared towards understanding and analysing the data, and exploiting in in a commercial sense. Typical Data Warehouses use OLAP technologies to find statistical regularities in data sets and try to construct prediction rules.

Entities, Attributes, Relationships:

These refer to the *Entity-Relationship* model, used since the 70s' to model relational databases. This model professes that any object in the world can be abstracted by either an entity (if it is an independent object) or a relationship (if it only exists through other objects). Attributes are the atomic characteristics of entities or relationships. This model can be applied to both relational or XML databases.

Primary Key:

A primary key is a set of attributes that identify in a unique manner an entity. In the real world, the couple FirstName/LastName is *not* a primary key of the entity [person], since there can exist two different people that have the same firstname and lastname. However, in the context of people in the W3C arenas, this seems to be the case.

Schema:

The schema of a Database is a description of the information it contains. In the case of Relational databases, the information contained is very constrained, for instance if you define a database with a [person] that contains attributes @firstname and @lastname, they are compulsory. In the case of XML Schemas, these are by far more flexible and therefore are much easier to change if one wishes to change the nature of the information already in the database, while not changing the queries that already run on the database.

Table:

A table is the atomic entity of a relational database in which it stores information. Typically, an entity can be seen as a table, and each of its attributes is transformed into a column of this table. Each line in the table represents an entry in the database.

Tree Structure:

A tree structure is the inherent structure of a mailing list discussion: it has one root node (the first message posted in the discussion) and has other nodes that branch out of this, that represent the answers to this initial message. In turn, each message can have answers. The result is called a *tree* in graph theory. The XML Model is based on trees, as opposed to the Relational model based on tables. One main difference can be seen when asking the query "find all the messages that answer directly or via another message to an initial message". This query is simple in XML, but impossible in SQL.

Web:

The World Wide Web, WWW or simply Web, was invented by Tim Berners-Lee in 1989 at Cern. The Web is a large pool of digital information available to anyone that can connect to it. The Web is governed by standards, defined by the World Wide Web Consortium, that was created in 1994. The most know standard is HTML, the format in which web pages are written.

REFERENCES

ABBOTT A. (1995) "Sequence Analysis", *Annual Review of Sociology*, 21:93-113.

ABBOTT A., BARMAN E. (1997) "Sequence Comparison via Alignment and Gibbs Sampling", *Sociological Methodology*, 27: 47-87.

ABBOTT A. GILBERT N. (2005), "Introduction", *American Journal of Sociology*, Volume 110 Number 4 (January): 859–863.

ABBOTT A., HRYCAK A. (1990) "Measuring Resemblance in Social Sequences", *American Journal of Sociology*, 96:144-185.

ABBOTT A., TSAY A. (2000) "A Sequence Analysis and Optimal Matching Methods in Sociology", *Sociological Methods and Research*, Vol 29, n°1, 3-33

ABITEBOUL S. (1997) "Querying Semi-Structured Data", in *Proceedings of the International Conference on Database Theory*.

ABITEBOUL S., COBENA G., NGUYEN B., POGGI A. (2002) "Sets of Pages of Interest". In *Bases de Données Avancées*.

ABITEBOUL S. (2003) "Managing an XML Warehouse in a P2P Context" In CAiSE Conference

AGUILERA V., BOISCUVIER F., CLUET S., KOECHLIN B. (2002), "Pattern tree matching for XML queries", Gemo Technical Report number 211, Available at http://www-rocqinria fr/gemo/Publication

AURAY N., CONEIN B., DORAT R., LATAPY M. (2007) "Multi-level analysis of an interaction network between individuals in a mailing-list", *Annals of Telecommunications*, Vol. 62, n°3-4, March-April.

BEAUDOUIN V., VELKOVSKA J. (1999) "Constitution d'un espace de communication sur internet", *Réseaux*, n° 97, 123-177.

BEAUDOUIN V., FLEURY S., PASQUIER M., HABERT B., LICOPPE C. (1999) "Décrire la toile pour mieux comprendre les parcours", *Réseaux*, n° 116 , 19-51

BENZECRI J. P. ET AL (1973) *L'Analyse des données*, Paris, Dunod.

BERKOWITZ S.D. (1982) *An Introduction to structural analysis*, Toronto Butterworth.

BRAGA D., CAMPI A., CERI S. (2004) "XQBE: A Graphical Interface for XQuery Engines", EDBT Conference: 848-850.

BREIGER R.L., BOORMAN S.A., ARABIE P. (1975) "An Algorithm for Clustering Relational Data with Application to social Network Analysis and Compariason with Multidimensional Scaling", *Journal of Mathematical Psychology*, 12

BRUNSSON N., JACOBSSON B. (2002) *A World of Standards*, Oxford University Press.

BUCKNER K., GILLHAM M. (1999) "Using E-Mail for Social and Domestic Purposes", *IFIP Conference Proceedings*, Vol. 173.

CAREN N., PANOFSKI A. (2005) "TQCA", *Sociological Methods and Research* vol 34, n°2, 147-172.

CARLEY K.M. (1996) "Artificial intelligence within sociology", *Sociological Methods and Research*, Vol. 25, n°1, 3-30.

CHAMBERLIN D., ROBIE J., AND FLORESCU D. (2000) "Quilt: An XML query language for heterogeneous data sources", *Proceedings of the International WebDB workshop*, Houston, USA.

CHAMBERLIN D. (2003) "XQuery: A query language for XML" In SIGMOD, p. 682. Slides available at http://www almaden ibm com/cs/people/chamberlin

CHATEAURAYNAUD F. (2003) "Marlowe - Vers un générateur d'expériences de pensée sur des dossiers complexes", Bulletin de Méthodologie Sociologique, n° 79, juillet 2003, Available at http://propsero.dyndns.org:9673/prospero/acces_public/06_association_doxa/BMS_MRLW

CHAUDHURI S., DAYAL U. (1997) "An overview of Data Warehousing and OLAP Technology" SIGMOD Record.

CODD, E. F. (1970) « A relational model of data for large shared data banks », in *Communications of the ACM, 13(6):377–387.*

DEUTSCH A., FERNANDEZ M., FLORESCU D., LEVY A., SUCIU D. (1999) "A query language for XML", *Proceedings of the International WWW Conference,* volume 31(11-16), 1155-1169.

DOREIAN P. (2001) "Causality in social network analysis", *Sociological Methods and Research*, vol 30, n°1, 81-114.

Dudouet F.-X. (2003) « De la régulation à la répression des drogues. Une politique publique internationale » in *Les Cahiers de la sécurité intérieure*, N°52, 2e trimestre.

DUDOUET F.-X., MANOLESCU I., NGUYEN B. SENELLART P., "*XML Warehousing Meets Sociology*", *Proceedings of the IADIS International Conference on the Web and Internet*, Lisbon, Portugal, October 2005

DUDOUET F.-X., MERCIER D., VION A. (2006) "Politiques de normalisation. Jalons pour la recherche empirique", *Revue Française de science politique*, vol 56, n° 3, June, 367-392.

FERNANDEZ M. (2004) "The Statesman, The General, His Lie tenant, and Her Sentry", *Keynote speech at the 1st International Workshop on Xquery Implementation Experience and Perspectives* (XIME-P)

GOODMAN N. (1978) *Ways of Worldmaking*, Hackett Publishing Co, Indianapolis.

GRAZ J.C. (2006) "Les hybrides de la mondialisation", *Revue Française de Science Politique*, vol 56, n° 6, November.

HAAS L., KOSSMANN D., WIMMERS E., YANG J. (1997) "Optimizing Queries Across Diverse Data Sources". VLDB 50, 276-285.

HOX J.J., KREFT I.G.G. (1994) "Multilevel analysis methods", *Sociological Methods and Research*, vol 22, n°3, 283-299.

HUISMAN M., SNIJDERS T. (2002) "Statistical Analysis of Longitudinal Network Data with changing composition", *Sociological Methods and Research*, Vol 30, n° 2, 425-454.

JANSEN I., VAN DEN TROOST A., MOLENBERGHS G., VERMULST A.A., GERRIS J.R.M. (2006) "Modeling Partially Incomplete Marital Satisfaction Data", *Sociological Methods & Research, 8,vol. 35: pp. 113 - 136.*

KOSSINETS G. (2006) "Effects of missing data in social networks", *Social Networks*, Volume 28, Issue 3, 1 July 2006, 247-268.

KALASHNIKOV D., CHEN S., NURAY R., MEHROTRA S., ASHISH, N. (2007) "Disambiguation Algorithm for People Search on the Web", *Proceedings of the 23rd IEEE International Conference on Data Engineering*, April 2007.

KRAAIJ W, WESTERVELD T., HIEMSTRA D. (2002) "The Importance of Prior Probabilities for Entry Page Search", *25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, August 11-15, Tampere, Finland.

LESNARD L., SAINT-POL T. (de) (2004) *Introduction aux méthodes d'appariement optimal (Optimal Matching Analysis)*, Document de travail INSEE, n° 2004-15, 30 p.

MARSDEN P.V., FRIEDKIN N.E. (1993) "Network studies of social influence", *Sociological Methods and Research*, vol 22, n°1, 127-151.

MARTIN C., VION A (2001) *Lone parent families and social care. A qualitative comparison of care arrangements in Finland, Italy, Portugal, United Kingdom and France*, Report to the EU Commission, october, 90 p. Available at www.uta.fi/laitokset/sospol/soccare/reports.htm

POWELL W.W., WHITE D.R., KOPUT K.W., SMITH J.O. (2006), "Growth of interorganizational collaboration the Life Sciences", *American Journal of Sociology*, Volume 111 Number 5 (March): 1367–1411

SAHUGUET A. (2000) The Kweelt system. Available at http://sourceforge net/projects/kweelt

STOVEL K, SAVAGE M, BEARMAN P (1996) « Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890-1970 », *American Journal of Sociology*, September, 358-399.

STARK D., VEDRES B. (2006) "Social Times of Network Spaces: Network Sequences and Foreign Investment in Hungary", *American Journal of Sociology*, Volume 111, Number 5 (March): .

TAMM-HALLSTRÖM K. (2001) "In Quest of Authority and Power: Standardization Organizations at Work", Scancor Workshop: Transnational regulation and the transformation of states California, USA 22-23 June.

TAMM-HALLSTRÖM K (2004) *Organizing International Standardization – ISO and the IASC in Quest of Authority*, Cheltenham United Kingdom 2004.

TOMASIC A., RASCHID L., VALDURIEZ P. (1996) "Scaling Heterogeneous Databases and the Design of Disco". *ICDCS*, 449-457

VAISMAN A.A. "OLAP, Data Warehousing, and Materialized Views: A Survey. " Available at: citeseer nj nec com/vaisman98olap html

WANG F., ZANIOLO C., ZHOU X (2005) "Temporal XML? SQL Is Fighting Back!" 12th International Symposium on Temporal Representation and Reasoning (TIME'05), June, 47-55.

WIDOM J. (1995) "Research problems in Data Warehousing", *International Conference on Information and Knowledge Management*.

WIEDERHOLD G (1992) "Mediators in the Architecture of Future Information Systems" *IEEE Computer* 25(3): 38-49

WHITE R., JOSE J.J., RUTHVEN I. (2001) "Query-Based Web Page Summarisation: A Task-Oriented Evaluation", 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, New Orleans, Louisiana, USA.

ZHAO S. (2006), "Humanoid social robot as a medium of communication", *New Media Society*, June 1, 8 (3), 401-419.

WEBSITES

QizX Open: a free-source Xquery Engine. Available at http://www.axyana.com/qizxopen/

Active XML reference: http://www.axml.net/

The DBWorld mailing list Available at http://www.cs.wisc.edu/dbworld

Projet e dot: http://www-rocq.inria.fr/gemo/Projects/edot/

IEEE Standardization Working Groups Areas Available at http://grouper ieee org/groups/index html

The KelKoo comparative shopping engine Available at http://www.kelkoo.com

The Linux Kernel mailing list archive Available at http://www.uwsg.indiana edu/hypermail/linux/kernel

The W3C Math Home Page Available at http://www.w3.org/Math

The Web Content Accessibility Guidelines Working Group Available at http://www .w3.org/WAI/GL

XML Path Language Available at http://www.w3.org/TR/xpath

The W3C XQuery mailing list (access restricted to W3C members) Available at http://lists.w3.org/Archives/Member/w3c-xml-query-wg

XQuery products and prototypes Available at http://www.w3 .org/XML/Query#Products

The XQL query language Available at http://www.w3.org/TandS/QL/QL98/pp/xql.html

The W3C XQuery Working Group Available at http://www.w3.org/XML/Query

TDA is available at http://steinhaus.stat.ruhr-uni-bochum.de/binaries.html, the interface Win TDA at: http://www.tufts.edu/~kschmi04/research/ and the manual at http://www.stat.ruhr-uni-bochum.de/tman.html

The Extensible Stylesheet Language Family Available at http://www.w3. org/Style/XSL

Available at http://search.cpan.org/~simon/Mail-Thread

Action Concertée Incitative Normes Pratiques et Régulations des Politiques Publiques Available at: http://www-rocq.inria.fr/gemo/Gemo/Projects/npp/index

XML Spy Available at: www altova com [52] XS m: the XML Summary Drawer Available at http://www-rocq.inria.fr/gemo/Gemo/Projects/SUMMARY

Zawinski J Message threading Available at http://www.jwz.org/doc/threading.html